# IMPLEMENTASI BIG DATA MENGGUNAKAN MODEL PEMROGRAMAN MAPREDUCE DAN KMEANS CLUSTERING

Darsin
Institut Teknologi Bisnis dan Bahasa Dian Cipta Cendikia
*Jl. Negara No. 03 Candi Mas Kotabumi Lampung Utara– Lampung*
*Email: itbadcc.darsin@gmail.com,*

***ABSRACT***

*Big data refers to very large and complex data sets, which are difficult to process using traditional methods of data processing. This data can come from various sources, such as social media, IoT (Internet of Things) devices, online transactions, and more. This very rapid development of data makes organizations look for methods to store and process data. Big data technology is a solution for storing data and also being able to process that data. The use of MapReduce is very important in big data processing, especially when the data that must be processed is very large and spread across many machines or servers. MapReduce is a programming model used to process and produce output from large amounts of data in a parallel and distributed manner. MapReduce allows dividing large tasks into small parts that can be processed on many machines or nodes. Hadoop is a framework that can store data on a large scale without paying attention to the structure of the data. To handle large amounts of data or big data, various technologies and techniques are needed that are more sophisticated than traditional data processing. The main aim of this research is to harness the enormous potential of existing data to create deeper insights, improve operational efficiency, and support smarter decision making. Another goal is also to propose and implement a Hadoop MapReduce based system. By utilizing the power of Hadoop MapReduce, it is hoped that the implementation of big data will have various significant benefits in many sectors so that it can ultimately encourage better performance and success in various aspects of business or operations. The results and conclusions from research on implementing K-Means in a big data environment using the MapReduce programming model are that the implementation of K-Means on liver disorder data using the Mahout library can run correctly. K-Means computation using the Mahout library produces the same output of centroid data items as is also proven by manual calculations.*

*Keywords: Big Data, Programming, Mapreduce and Kmeans Clustering*

**ABSTRAK**

Big data merujuk pada kumpulan data yang sangat besar dan kompleks, yang sulit untuk diolah dengan menggunakan metode tradisional dalam pengolahan data. Data ini dapat datang dari berbagai sumber, seperti media sosial, perangkat IoT (Internet of Things), transaksi online, dan banyak lagi. Perkembangan data yang sangat pesat ini membuat organisasi mencari metode untuk menyimpan dan mengolah data. Teknologi big data menjadi solusi untuk menyimpan data dan juga mampu mengolah data tersebut. Penggunaan MapReduce sangat penting dalam pemrosesan big data, terutama ketika data yang harus diproses sangat besar dan tersebar di banyak mesin atau server. MapReduce merupakan model pemrograman yang digunakan untuk memproses dan menghasilkan output dari data dalam jumlah besar secara paralel dan terdistribusi. MapReduce memungkinkan pembagian tugas besar menjadi bagian-bagian kecil yang bisa diproses di banyak mesin atau node. Hadoop merupakan sebuah framework yang dapat menyimpan data dalam skala besar tanpa memperhatikan struktur dari data untuk menangani data yang banyak atau big data, diperlukan berbagai teknologi dan teknik yang lebih canggih dibandingkan dengan pengolahan data tradisional. Penelitian ini tujuan utamanya adalah untuk memanfaatkan potensi besar data yang ada untuk menciptakan wawasan yang lebih dalam, meningkatkan efisiensi operasional, dan mendukung pengambilan keputusan yang lebih cerdas. Tujuan lain juga untuk mengusulkan dan

menerapkan sebuah sistem berbasis Hadoop MapReduce. Dengan memanfaatkan kekuatan Hadoop MapReduce diharapkan implementasi big data memiliki berbagai manfaat yang signifikan di banyak sektor sehinnga akhirnya dapat mendorong kinerja dan keberhasilan yang lebih baik dalam berbagai aspek bisnis atau operasional. Hasil dan kesimpulan dari penelitian implementasi K-Means dalam lingkungan big data menggunakan model pemrograman MapReduce adalah Implementasi K-Means pada data liver disorder menggunakan library Mahout dapat berjalan dengan benar. Komputasi K-Means dengan menggunakan library Mahout menghasilkan output item data centroid yang sama karena dibuktikan juga dengan penghitungan manual.

Kata Kunci : Big Data, Pemrograman, Mapreduce dan Kmeans Clustering

## 1. PENDAHULUAN

Pada saat ini perkembangan teknologi yang sangat besar telah memunculkan sebuah sistem big data sebagai tempat penyimpanan segara informasi (Wardani et al., 2023). Big data sendiri adalah sekumpulan data dengan jumlah yang sangat besar yang memiliki kemampuan untuk mengumpulkan, menganalisa, dan menyimpan data yang sangat banyak (Eni Kustanti, 2021) . Data- data yang tersimpan didalam big data sangat beragam sumbernya, diantaranya adalah media sosial, sensor, video surveillance, dan smart grids. Maka dari itu, big data sangat bermanfaat bagi pihak-pihak yang membutuhkan (Nurul Fajriyah1 & , Wawan Setiawan2, Ernawati Dewi3, 2022).

Big data merujuk pada kumpulan data yang sangat besar dan kompleks, yang sulit untuk diolah dengan menggunakan metode tradisional dalam pengolahan data (Saudjhana et al., 2024). Data ini dapat datang dari berbagai sumber, seperti media sosial, perangkat IoT (Internet of Things), transaksi online, dan banyak lagi (Sedayu & Andriyansah, 2021). Perkembangan data yang sangat pesat ini membuat organisasi mencari metode untuk menyimpan dan mengolah data. Teknologi big data menjadi solusi untuk menyimpan data dan juga mampu mengolah data tersebut (Beni Ilham Priyambodo, 2023). Hadoop merupakan sebuah framework yang dapat menyimpan data dalam skala besar tanpa memperhatikan struktur dari data Untuk menangani data yang banyak atau big data, diperlukan berbagai teknologi dan teknik yang lebih canggih dibandingkan dengan pengolahan data tradisional (Arrahmando Romadhona et al., 2022).

Penggunaan MapReduce sangat penting dalam pemrosesan big data, terutama ketika data yang harus diproses sangat besar dan tersebar di banyak mesin atau server. MapReduce merupakan model pemrograman yang digunakan untuk memproses dan menghasilkan output dari data dalam jumlah besar secara paralel dan terdistribusi (Adawiyah & Munir, 2020). MapReduce memungkinkan pembagian tugas besar menjadi bagian-bagian kecil yang bisa diproses di banyak mesin atau node (Awad & Hamad, 2022). MapReduce banyak digunakan dalam ekosistem Hadoop dan juga dalam berbagai aplikasi lain yang memerlukan pemrosesan data dalam skala besar, seperti analisis data, pemrosesan log, atau pemrograman paralel (Ilhami et al., 2023).

Hadoop menggunakan konsep pemrograman MapReduce untuk mengolah data menjadi informasi (Awaluddin et al., 2023). Koleksi data yang besar dapat diolah dan dianalisis untuk mendapatkan nilai atatu *value* pada data. Hasil analisa data tersebut berupa informasi yang dapat dijadikan pengambilan kebijakan pada organisasi. MapReduce mampu melakukan komputasi secara paralel dan terdistribusi pada sistem Hadoop (Chunqiong Wu , 1, 2 Bingwen Yan , 1, 2 Rongrui Yu , 1, 2 Baoqin Yu, 1, 2 Xiukao Zhou, 1, 2 Yanliang Yu, 1, 2 and Na Chen2 & 1Business, 2021).

Algoritma K- Means ialah salah satu dari banyaknya proses yang dipakai dalam kegiatan cluster, metode inilah yang efektif untuk menghasilkan cluster-cluster dari data kecil maupun besar (Dikarya & Muharni, 2022). Data set merupakan data yang masih perlu diolah menjadi sebuah informasi dengan melakukan Pengelompokan data atau biasa disebut dengan data clustering. Metode clustering yang digunakan adalah K-Medoids dengan Mapreduce sebagai baseline penelitian (Putra et al., 2021). Kelebihan dari metode K-means ini mudah untuk diterapkan karena

menghitung nilai jarak centroid setiap cluster dengan menggunakan rumus Euclidience distance yang sangat umum digunakan (Lestari et al., 2022)

Penelitian ini tujuan utamanya adalah untuk memanfaatkan potensi besar data yang ada untuk menciptakan wawasan yang lebih dalam, meningkatkan efisiensi operasional, dan mendukung pengambilan keputusan yang lebih cerdas. Tujuan lain juga untuk mengusulkan dan menerapkan sebuah sistem berbasis Hadoop MapReduce (Venger & Akhtoian, 2021). Dengan menganalisis data yang besar dan beragam, sistem ini diharapkan dapat memberikan wawasan berharga bagi lembaga atau instansi dalam meningkatkan kinerjanya. Kami berharap penelitian ini dapat memberikan kontribusi yang berarti bagi perkembangan teknologi analisis big data. Dengan memanfaatkan kekuatan Hadoop MapReduce diharapkan implementasi big data memiliki berbagai manfaat yang signifikan di banyak sektor sehinnga akhirnya dapat mendorong kinerja dan keberhasilan yang lebih baik dalam berbagai aspek bisnis atau operasional.

## 2. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah kualitatif dengan pendekatan studi literatur. Studi literarur adalah kajian literatur dengan memakai pendekatan konseptual-tradisional (Solihin, 2021). Pada penelitian ini lebih merumuskan atau merancang dengan cara mencari sumber- sumber buku literatur yang sinkron dengan teori-teori yang diteliti, dan juga menganalisis artikel-artikel ilmiah. Seluruh artikel yang di sitasi bersumber dari Google Cendikia dan Mendele (Syira et al., 2023). Adapun tahapan penelitian akan dijabarkan sebagai berikut:

### 2.1. Studi Pustaka

Pada studI ini metodelogi yang digunakan adalah study literature review dari journal yang terkait dengan penelitian tentang Big Data (Siregar & Musawaris, 2023). Studi pustaka menjelaskan teori-teori yang digunakan dalam penelitian. Adapun teori-teori yang digunakan ialah *data mining*, K-Means *clustering*, *big data*, Hadoop, *Hadoop Distributed File System* (HDFS), Yarn, MapReduce, dan Mahout. Peneliti mengintegrasikan sudut pandang yang berbeda dari berbagai sumber literatur (Veri Ferdiansyah & Muhammad Irwan Padli Nasution, 2023)

### 2.2 Perancangan sistem

Perancangan sistem meliputi segala perangkat lunak dan perangkat keras yang dibutuhkan dalam mengembangkan sistem.

### 2.3 Luaran sistem

Luaran sistem ini ialah sebuah sistem *big data* dengan menggunakan menggunakan *framework* Hadoop. Sistem Hadoop ini berjalan pada jaringan lokal. *Library* Mahout yang berjalan pada sistem Hadoop digunakan untuk menganalisa koleksi data pada sistem Hadoop.

### 2.4 Evaluasi

Evaluasi sistem ini akan dibagi kedalam dua bagian yakni:

a. Membandingkan hasil komputasi K-Means dengan menggunakan *library* Mahout dengan menggunakan penghitungan manual. Hasil pengujian memperlihatkan kecocokan centroid dari hasil komputasi dengan menggunakan *library* Mahout berdasarkan hasil dari penghitungan manual.
b. Menguji unjuk kerja implementasi K-Means *clustering* pada lingkungan *big data.* Pengujian dilakukan dengan menjalankan komputasi K-Means menggunakan *library* Mahout sebanyak 10 kali pada jumlah *slave node* yang berbeda. Hasil pengujian memperlihatkan rata-rata waktu eksekusi komputasi K-Means pada jumlah *slave node* yang berbeda.

## 3 HASIL DAN PEMBAHASAN

Hasil dan pembahasan dari penelitian ini akan diuraikan lebih rinci dibawah ini.

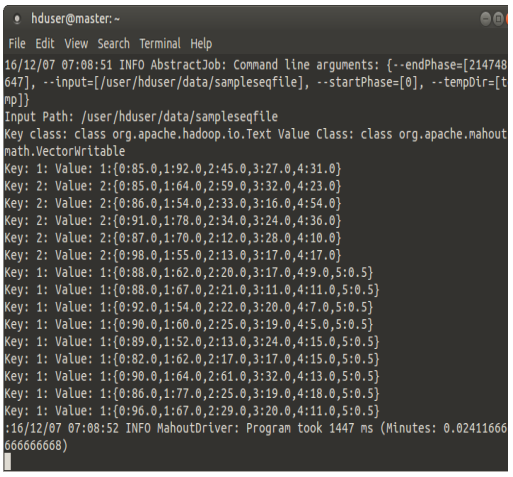
### 3.1 Implementasi Metode K-Means Menggunakan Library Mahout

Data *liver disorder* memiliki susunan informasi yakni id, mvc, alkphos, sqpt, sgot, gammagt, drink, dan selector. Namun dalam melakukan konversi, susunan file tersebut diubah menjadi

*P-ISSN: 2337-8344*
*E-ISSN: 2623-1247*

selector, id, mvc, alkphos, sqpt, sgot, gammagt, dan drink. Hasil atau *output* yang akan dicapai ialah file data *liver disorder* dan file centroid yang masing-masing berformat sequence. Adapun data centroid dipilih secara manual. Hasil *output* tersebut akan diinputkan ke dalam *Hadoop Distributed File System*.

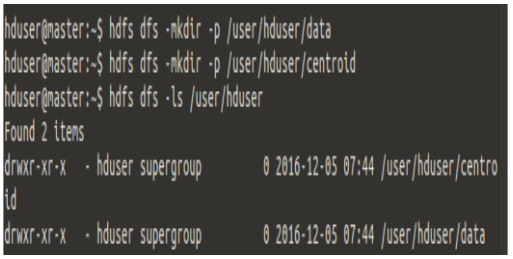
Adapun file *sequence* tidak *human readable* atau tidak dapat dibaca secara langsung. Gambar 1 menunjukkan fitur yang telah disediakan oleh Mahout untuk membaca *sequence file*. Adapun nama dari *sequence file* ialah sampleseqfile. Sehingga perintah yang digunakan ialah $mahout seqdumper -i sampeseqfile | less. Mahout seqdumper menjelaskan bahwa mahout menggunakan metode seqdumper. Parameter-i menjelaskan input file yang kemudian diikut olah nama file. | less merupakan perintah dari bash linux yang digunakan untuk menerima *output* dari perintah sebelumnya, lalu kemudian menginputkan perintah tersebut ke dalam perintah less. Less merupakan perintah bash yang digunakan untuk membaca sebuah file atau *output*.



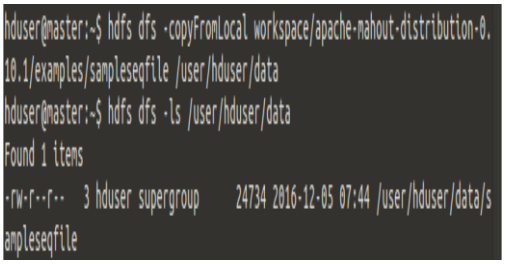

Gambar 1. Menjalankan metode seqdumper pada *command line*

Langkah selanjutnya ialah dengan menjalankan perintah start-dfs.sh dan start-yarn.sh, Perintah start-dfs.sh berfungsi untuk menjalankan *service*NameNode dan SecondaryNamenode pada *master node*, dan DataNode pada *slave node*. Sedangkan perintah start-yarn.sh berfungsi untuk menjalankan *service* ResourceManager pada *master node* dan *service* NodeManager pada *slave node*.

Gambar 2 menunjukkan proses membuat direktori pada HDFS. Perintah yang digunakan ialah $hdfs dfs -mkdir diikuti nama direkori. Direktori yang digunakan dalam penelitian ini ialah /user/hduser/data dan /user/hduser/centroid. Direktori /user/hduser/data digunakan untuk menyimpan data *training*. Sedangkan direktori /user/hduser/centroid digunakan untuk menyimpan data centroid. Parameter-p digunakan untuk mebuat direktori dalam direktori. Sehingga perintah yang digunakan ialah $hdfs dfs -mkdir -p diikuti nama direktori. Perintah $hdfs dfs -ls digunakan untuk melihat sebuah file atau direktori. Sehingga perintah dapat digunakan untuk mengecek keberhasilan dalam membuat direktori.

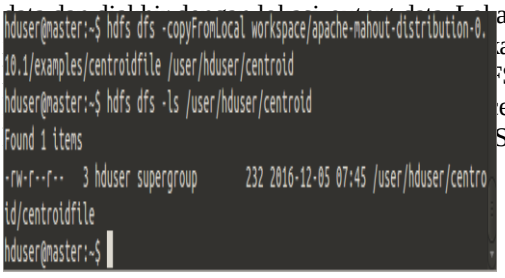



Gambar 2. Membuat direktori data dan direktori centroid pada hdfs




Gambar 3. menunjukkan proses menyimpan file data training pada HDFS.

Gambar 3 menyimpan sebuah file data *trining* dari sistem lokal ke dalam HDFS . Sedangkan gambar 4 menunjukkan proses menyimpan file data centroidpada HDFS. Perintah yang digunakan dalam menyimpan file ialah $hdfs dfs-copyFromLocal kemudian diikuti lokasi input



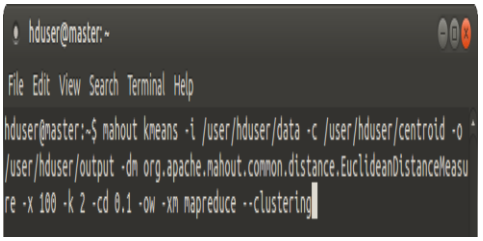
*P-ISSN: 2337-8344*
*E-ISSN: 2623-1247*

Gambar 4. Menyimpan sebuah file centroid pada sistem lokal ke dalam HDFS

### 3.2 Proses Menjalankan Komputasi K-Means

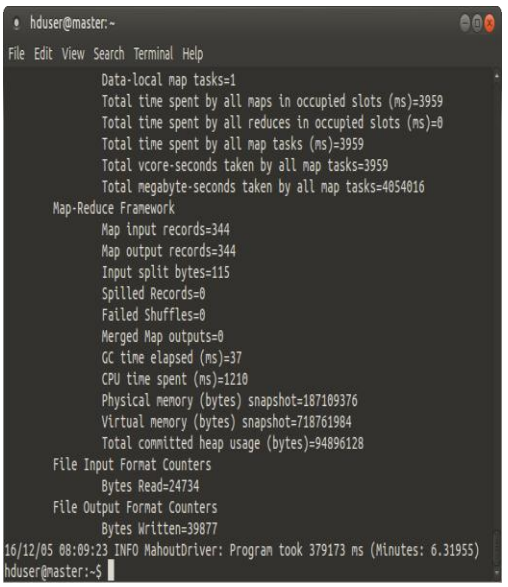
Perintah yang digunakan untuk menjalankan komputasi K-Means ialah $mahout kmeans diikuti lokasi file data, lokasi file centroid, lokasi file *output*, *distance measure* atau algoritma yang digunakan untuk menghitung jarang antara item data dan pusat *cluster* atau centroid, iterasi maksimal, jumlah K atau *cluster* dari data, *convergen delta* atau nilai untuk menentukan proses iterasi berhenti, *execution method* atau metode yang digunakan untuk mengeksekusi data, dan *clustering* menentukan agar proses *clustering* berjalan setelah proses iterasi telah berlangsung.

Adapun semua lokasi data yang digunakan dalam perintah $mahout kmeans berada pada HDFS. Sehingga lokasi *output* data juga terdapat dalam HDFS. Gambar 5 menunjukkan bahwa *Distance Measure* (parameter -dm) yang digunakan dalam penelitian ini ialah Euclidean *Distance Measure*. Iterasi maksimal (parameter -x) yang digunakan berjumlah 100. Total K (parameter -k) yang digunakan berjumlah 2. Hal ini dikarenakan data dikelompokkan dalam 2 kategori yakni kelompok yang memiliki kelainan hati (*liver disorder)* dan kelompok yang tidak memiliki kelainan hati (*non liver disorder*). *Convergen delta* (parameter -cd) yang digunakan bernilai 0.1. *Execution method* (parameter -xm) yang digunakan ialah mapreduce. Parameter -ow atau –overwrite berfungsi untukmemaksa sistem untuk menulis hasil *output* walaupun sistem sudah memiliki lokasi *output* tersebut sebelumnya.



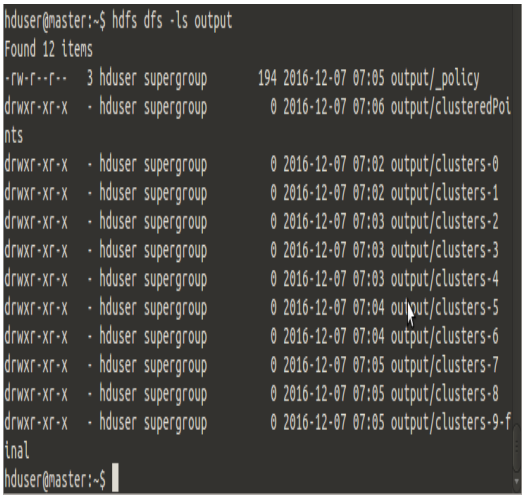

Gambar 5. Menjalankan K-Means menggunakan *library* Mahout

Gambar 6 menunjukkan hasil akhir dari iterasi yan gditunjukkan dengan menampilkan beberapa metadata dari proses K-Means dan waktu komputasi dijalankan.




Gambar 6. Akhir dari iterasi K-Means

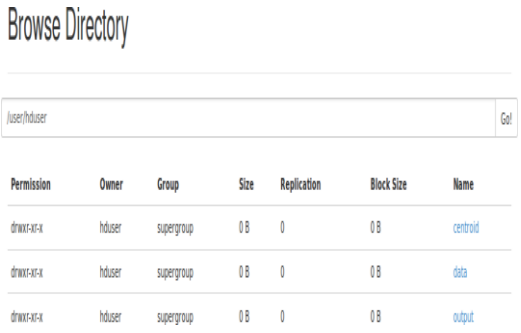
Gambar 7 menunjukkan perintah $hdfs dfs -ls dapat digunakan untuk melihat hasil dari proses K-Means. Sedangkan perintah $hdfs dfs -ls *output* atau $hdfs dfs -ls /user/hduser/*output* digunakan untuk melihat isi dari direktori *output.* Hasil proses K-Means dapat dilihat dari aplikasi NameNode *Web Interface y*ang beralamat di master:50070.




Gambar 7. Perintah $hdfs dfs -ls *output*

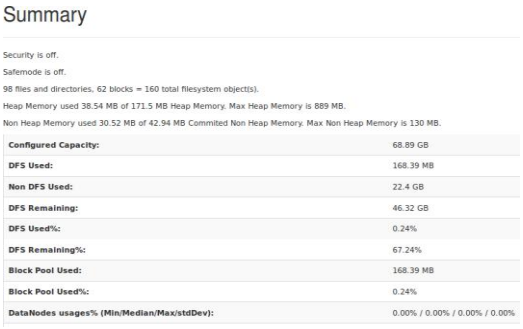
Gambar 8 menunjukkan hasil K-Means melalui aplikasi NameNode *web interface*.

*P-ISSN: 2337-8344*
*E-ISSN: 2623-1247*



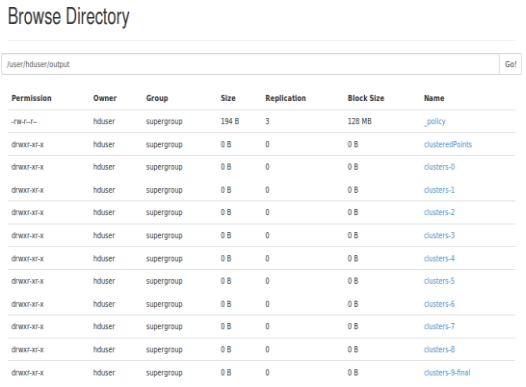

Gambar 8. Hasil proses K-Means pada direktori /user/hduser/*output*

Gambar 9 menunjukkan bahwa aplikasi ini dapat menunjukkan aktivitasdan semua informasi dari sistem Hadoop baik sebelum dan sesudah komputasi MapReduce dijalankan. Beberapa informasi yang ditunjukkan pada sistem ini ialah DataNode *information* atau informasi DataNode (*slave node*). Informasi DataNode ini menunjukkan slave *node* yang aktif dan tidak aktif, kapasitas HDFS dari masing-masing DataNode, dan *storage* HDFS yang telah digunakan, Aplikasi ini juga dapat digunakan untuk mencari dan melihat data pada HDFS.



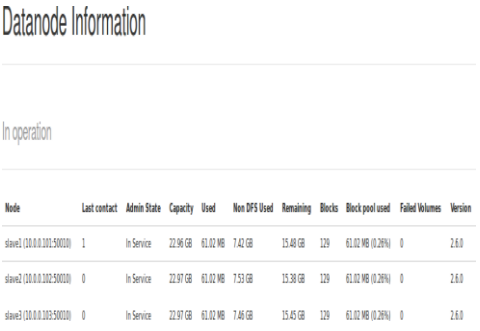

Gambar 9. Summary berisi ringkasan informasi *DataNode*

NameNode *web interface* juga dapat memperlihatkan informasi tentang DataNode. Gambar 10 menampilkan beberapa informasi DataNode pada sistem Hadoop. Beberapa informasi DataNode yang ditampilkan ialah alamat *slave node*. Status *slave node* digunakan yang diperlihatkan pada Admin State. Informasi ukuran HDFS juga diperlihatkan seperti total ukuran disk yang dapat digunakan untuk menyimpan data (*capacity*) dan total kapasitas disk yang telah digunakan (*used*).




Gambar 10. Informasi DataNode pada aplikasi NameNode *Web Interface*

Gambar 11 menjelaskan bahwa HDFS memiliki direktori data, centroid, dan *output*.



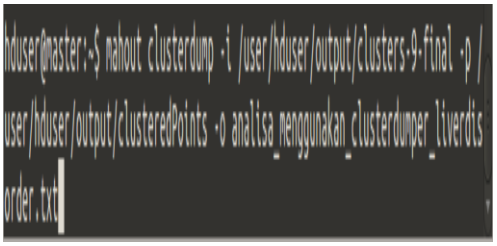

Gambar 11. Informasi direktori /user/hduser pada HDFS

Gambar 12 menunjukkan informasi direktori /user/hduser/data pada HDFS yakni *permission, owner, group, size, replication, block size,* dan *name*. *Permission* merupakan hak akses atau *priviledge* yang diijinkan olah sistem Hadoop. *Owner* merupakan pemilik atau pengguna sistem Hadoop. *Group* merupakan pengelompokkan data pada sistem Hadoop. *Replication* menjelaskan bahwa data mengalami proses replikasi di beberapa *node*. *Block size* merupakan ukuran blok data. Sedangkan *name* merupakan name file tersebut.

*P-ISSN: 2337-8344*
*E-ISSN: 2623-1247*


Gambar 12. Informasi direktori /user/hduser/data

Aplikasi ini juga mampu menampilkan metadata dari file pada HDFS. Gambar 13 menampilkan *metadata* dan *availability* dari file sampleseqfile. Availability menunjukkan bahwa file sampleseqfile telah mengalami proses replikasi pada 3 *DataNode*.




Gambar 13. Informasi file sampleseq file

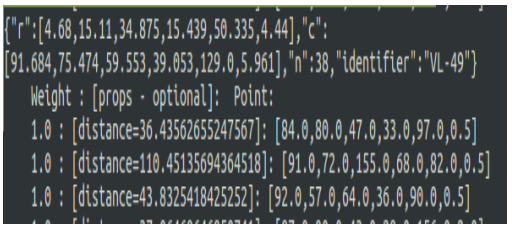
### 3.3 Analisa Implementasi K-Means Menggunakan Library Mahout Pada Lingkungan Big Data

Hasil yang diperoleh dari algoritma K-Means ialah berupa direktori /user/hduser /*output* pada HDFS. Mahout menyediakan metode untuk menganalisa hasil dari komputasi -Means. Metode yang digunakan ialah dengan menggunakan perintah clusterdump. Metode ini dapat membuat atau *generate* file analisa yang mengelompokkan *item* data berdasarkan centroid atau pusat kelompoknya. Perintah yang digunakan ialah $mahout clusterdump -i direktori iterasi terakhir -p clusteredPoints -o *output*_file_ analisa. Parameter -i atau –input merupakan perintah untuk memeasukkan direktori input yang berupa direktori iterasi terakhir. Parameter-p atau –pointsDir merupakan perintah untukmenginputkan direktori clusteredPoints yang berupa hasil akhir dari data yang telah mengalami proses komputasi K-Means. Sedangkan parameter -o atau – *output* digunakan untuk menghasilkan file analisa *output*.

Gambar 14 menunjukkan direktori iterasi terakhir adalah user/hduser/o*tput*/clusters-9-final. Direktori clusters-9-final memiliki data centroid akhir. Direktori /user/hduser/ *output*/clusteredPoints terdiri dari data hasil iterasi. Sedangkan analisa_menggunakan_ clusterdumper_liverdisorder.txt merupakan file hasil analisa.


Gambar 14: Analisa hasil K-Means menggunakan clusterdump

Hasil dari clusterdump ditampilkan pada gambar 2. VL-27 merupakan identitas yang secara otomatis diberikan oleh Mahout. “n” merupakan total data pada kelompok tersebut. “c” merupakan centroid atau pusat kelompok akhir. “r” merupakan radius dari kelompok. Sedangkan dibaris selanjutnya merupakan pengelompokkan data yang telah masuk kategori kelompok VL-27.

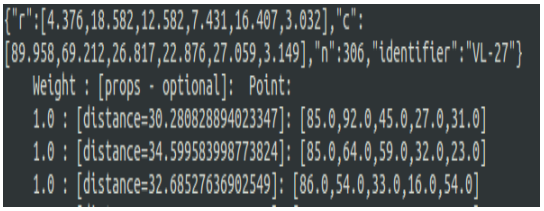



Gambar 15. Beberapa hasil analisa *cluster* data dengan identitas VL-27

Gambar 15 menunjukkan bahwa kelompok VL-27 memiliki total *item* data sebanyak 306. Sedangkan pada gambar 3 menunjukkan bahwa kelompok VL-49 memiliki total *item* data sebanyak 38. Sehingga total *item* data keseluruhan

*P-ISSN: 2337-8344*
*E-ISSN: 2623-1247*

ialah 344.




Gambar 16. Beberapa hasil analisa *cluster* data dengan identitas VL-49

Pengujian hasil implementasi K-Means menggunakan *library* Mahout dilakukan dengan membandingkan dengan hasil dari penghitungan manual. Penghitungan manual juga menggunakan data centroid yang sama dengan penghitungan menggunakan *library* Mahout. Penghitungan manual menggunakanlabel centroid C1 dan C2. Item data pada centroid C1 berjumlah 306, sedangkan item data pada centroid C2 berjumlah 38.

Tabel 1 menunjukkan bahwa hasil penghitungan manual memiliki item data centroid yang sama dengan hasil penghitungan menggunakan *library* Mahout. Sehingga dapat disimpulkan bahwa *library* Mahout dapat melakukan komputasi K-Means dengan benar.

**Tabel 1**. Perbandingan hasil penghitungan manual dan library Mahout

| Label Centroid | Item Data Centroid | | | | | | Jumlah Item Data |
|---|---|---|---|---|---|---|---|
| C1 | 8.995.751 634 | 6.921.241 83 | 268.169 9346 | 2.287.581 699 | 2.705.882 353 | 3.148.692 81 | 306 |
| VL-27 | 89.958 | 69.212 | 26.817 | 22.876 | 27.059 | 3.149 | 306 |
| C2 | 9.168.421 53 | 7.547.368 421 | 595.526 3158 | 3.905.263 158 | 129 | 5.960.526 316 | 38 |
| VL-49 | 91.684 | 75.474 | 59.553 | 39.053 | 129 | 5.961 | 38 |

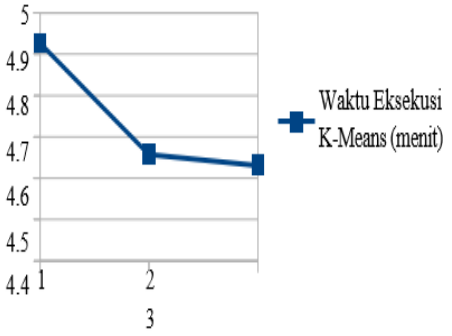
### 3.4 Analisa Unjuk Kerja Implementasi K-Means Menggunakan Library Mahout Pada Lingkungan Big Data

Analisa unjuk kerja dilakukan dengan menjalankan komputasi K-Means menggunakan *library* Mahout sebanyak 10 kali dalam jumlah slave *node* yang berbda. Nilai rata-rata digunakan untuk mengevaluasi hasil unjuk kerja. Tabel 2 menunjukkan bahwa semakin banyak jumlah slave *node* maka waktu eksekusi K- Means semakin cepat.

Tabel 2**.** Unjuk kerja implementasi K-

| Percobaan ke- | Waktu eksekusi K-Means pada Slave Node (Menit) | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 4.523 | 4.646 | 02.08 |
| 2 | 5.295 | 5.029 | 4.553 |
| 3 | 4.525 | 0,2319444 | 4.946 |
| 4 | 0,2333333 | 0,2326389 | 5.461 |
| 5 | 0,23125 | 4.058 | 2.759 |
| 6 | 5.397 | 5.384 | 4.967 |
| 7 | 4.001 | 5.368 | 4.544 |
| 8 | 4.923 | 3.631 | 5.449 |
| 9 | 5.297 | 4.045 | 5.434 |
| 10 | 5.428 | 4.548 | 5.416 |
| Rata-rata | 49.279 | 46.599 | 46.329 |

Gambar 17 menunjukkan semakin banyak jumlah *slave node* yang digunakan maka waktu eksekusi untuk menjalankan K-Means menggunakan *library* Mahout semakin cepat.




Gambar 17. Grafik hasil unjuk kerja sistem Hadoop

## 4 KESIMPULAN DAN SARAN

Kesimpulan penelitian implementasi K-Means dalam lingkungan *big data* menggunakan model pemrograman MapReduce adalah Implementasi K-Means pada data *liver disorder* menggunakan *library* Mahout dapat berjalan dengan benar. Komputasi K-Means dengan menggunakan *library* Mahout menghasilkan *output item* data centroid yang sama dengan karean dibuktikan juga penghitungan manual. Hasil unjuk kerja menunjukkan bahwa semakin banyak *slave*

*node* yang digunakan maka semakin cepat waktu yang dibutuhkan untuk menjalankan komputasi K-Means yang menggunakan *library* Mahout. Saran dari hasil penelitian implementasi K-Means pada lingkungan *big data* ini, penelitian selanjutnya dengan topik yang sama ialah menggunakan algoritma data mining yang berbeda.

**DAFTAR PUSTAKA**

[1] Adawiyah, R., & Munir, S. (2020). Jurnal Informatika Terpadu ANALISIS DAN EVALUASI ALGORITMA MAPREDUCE WORDCOUNT PADA CLUSTER HADOOP MENGGUNAKAN INDIKATOR KECEPATAN. *Jurnal Informatika Terpadu*, *6*(1), 14–19. https://journal.nurulfikri.ac.id/index.php/JIT

[2] Arrahmando Romadhona, L., Febrianti, R., Winata, A., Putri Amanda, C., & Julia Erizka, R. (2022). NetPLG Journal of Network and Computer Applications Hadoop-MapReduce Pada YARN Framework. *Journal of Network and Computer Applications*, *1*(2), 91–101. https://jurnal.netplg.com/jnca

[3] Awad, F. H., & Hamad, M. M. (2022). Improved k-Means Clustering Algorithm for Big Data Based on Distributed SmartphoneNeural Engine Processor. *Electronics (Switzerland)*, *11*(6). https://doi.org/10.3390/electronics11060883

[4] Awaluddin, M., Angelia Mahlil, R., & Ode Muhammad Saidi, L. (2023). Implementasi Hadoop Mapreduce Untuk Memprediksi Predikat Kelulusan Mahasiswa. *Journal on Education*, *05*(04), 17239–17251.

[5] Beni Ilham Priyambodo. (2023). PEMANFAATAN BIG DATA UNTUK PENINGKATAN BISNIS BANK. *Jurnal Ilmiah Indonesia*, *VIII*(10), 1–19. https://doi.org/http://dx.doi.org/10.36418/syntax-literate.v6i6

[6] Chunqiong Wu , 1, 2 Bingwen Yan , 1, 2 Rongrui Yu , 1, 2 Baoqin Yu, 1, 2 Xiukao Zhou, 1, 2 Yanliang Yu, 1, 2 and Na Chen2, 3, & 1Business. (2021). *k-Means Clustering Algorithm and Its Simulation Based on Distributed Computing Platform* (pp. 1–10). https://doi.org/https://doi.org/10.1155/2021/9446653

[7] Dikarya, F., & Muharni, S. (2022). Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Universitas Terbaik Di Dunia. *Jurnal Informatika*, *22*(2), 124–131. https://doi.org/10.30873/ji.v22i2.3324

[8] Eni Kustanti. (2021). Implementasi Big Data Pada Manajemen Pengetahuan Komoditas Pertanian. *Jurnal Pustakawan Indonesia*, *20*(1), 1–9.

[9] Ilhami, A. R., Putra, A., Aurelie, A. C., Junaidi, K., & Israwana, S. (2023). Implementasi MapReduce Pada Dataset Spotify Top Music Untuk Mengetahui Artist yang paling banyak Didengar Dalam Kurun Waktu 10 Tahun. *Journal of Network and Computer Applications*, *2*(2), 37–51. https://jurnal.netplg.com/

[10] Lestari, D., Charis Fauzan, A., Ilmu Eksakta, F., Studi Ilmu Komputer, P., Nahdlatul Ulama Blitar, U., & Masjid No, J. (2022). Penerapan Algoritma Pillar Untuk Optimasi Penentuan Titik Awal Centroid Pada Algoritma K-Means Clustering. *JOISIE Journal Of Information System And Informatics Engineering*, *6*(1), 15–24.

[11] Nurul Fajriyah1, & , Wawan Setiawan2, Ernawati Dewi3, T. D. (2022). IMPLEMENTASI TEKNOLOGI BIG DATA DI ERA DIGITAL Nurul. *JURNAL INFORMATIKA*, *1*(1), 356–363.

[12] Putra, H. M., Akbar, T., Ahmadi, A., & Darmawan, M. I. (2021). Analisa Performa Klastering Data Besar pada Hadoop. *Infotek : Jurnal Informatika Dan Teknologi*, *4*(2), 174–183.

https://doi.org/10.29408/jit.v4i2.3565

[13] Saudjhana, A., Budiman, A., Fernando, H., Juliantio, J., Junianto, K., Venessa, K., Salim, S., & Tomy, T. (2024). Implementasi Big Data terhadap Pengecekan Medis dan Konsultasi Kesehatan di Indonesia. *Journal of Information System and Technology*, *5*(1), 1–6. https://doi.org/10.37253/joint.v5i1.4323

[14] Sedayu, A. S., & Andriyansah, A. (2021). Pemanfaatan Big Data pada Instansi Pelayanan Publik. *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, *4*(7), 543–548. https://doi.org/10.54371/jiip.v4i7.309

[15] Siregar, J. J., & Musawaris, R. (2023). Utilization of Big Data in the Education Sector. *Indonesian Journal of Social Technology*, *4*(2), 2745–5254. https://doi.org/Doi:10.59141/jist.v4i02.737

[16] Solihin, O. (2021). Implementasi Big Data Di Sosial Media Untuk Komunikasi Krisis Pemerintah. *Jurnal Common*, *5*(1), 56–66. https://doi.org/10.34010/common.v5i1.5123

[17] Syira, S. D., Fauzi, A., Woestho, C., Vilani, L., & ... (2023). Pemanfaatan Big Data dalam Peningkatan Efektivitas Strategi Komunikasi Marketing Terpadu pada Perusahaan E-Commerce. *Jurnal Ekonomi …*, *4*(5), 891–900. https://doi.org/https://doi.org/10.31933/jemsi.v4i5

[18] Venger, E. I., & Akhtoian, A. (2021). the Role of Big Data in the Implementation of Digital-Marketing Strategies. *Proceedings of Scientific Works of Cherkasy State Technological University Series Economic Sciences*, *12*(1), 61–68. https://doi.org/10.24025/2306-4420.63.2021.248464

[19] Veri Ferdiansyah, & Muhammad Irwan Padli Nasution. (2023). Penerapan Teknologi Big Data Dalam Pengembangan Database Pendidikan. *Jurnal Riset Manajemen*, *1*(3), 22–29. https://doi.org/10.54066/jurma.v1i3.591

[20] Wardani, S. D. K., Ariyanto, A. S., Umroh, M., & Rolliawati, D. (2023). Perbandingan Hasil Metode Clustering K-Means, Db Scanner & Hierarchical Untuk Analisa Segmentasi Pasar. *JIKO (Jurnal Informatika Dan Komputer)*, *7*(2), 191. https://doi.org/10.26798/jiko.v7i2.796